Large scale estimation of arterial traffic and structural analysis of traffic patterns using probe vehicles

Aude Hofleitner, Ryan Herring, Alexandre Bayen, Yufei Han, Fabien Moutarde, Arnaud de La Fortelle

Large scale estimation of arterial traffic and structural analysis of traffic patterns using probe vehicles

Aude Hofleitner* Ryan Herring† Alexandre Bayen‡
Yufei Han§ Fabien Moutarde§ Arnaud de La Fortelle§

91st Annual Meeting of the Transportation Research Board
January 22-26, 2012, Washington D.C

Word Count:

Number of words:	6811
Number of figures:	5 (250 words each)
Number of tables:	0 (250 words each)
Total:	8061

*Corresponding Author, Department of Electrical Engineering and Computer Science, University of California, Berkeley and UPE/IFSTTAR/GRETTIA, France, aude.hofleitner@polytechnique.edu

†Apple Inc. Affiliation during redaction of the paper: California Center for Innovative Transportation, Berkeley CA, ryanherring@berkeley.edu

‡Department of Electrical Engineering and Computer Science and Department of Civil and Environmental Engineering, Systems Engineering, University of California, Berkeley, bayen@berkeley.edu

§Robotics Lab (CAOR), Mines ParisTech, Paris, France, Yufei.Han@mines-paristech.fr
Fabien.Moutarde@mines-paristech.fr

Abstract

Estimating and analyzing traffic conditions on large arterial networks is an inherently difficult task. The first goal of this article is to demonstrate how arterial traffic conditions can be estimated using sparsely sampled GPS probe vehicle data provided by a small percentage of vehicles. Traffic signals, stop signs, and other flow inhibitors make estimating arterial traffic conditions significantly more difficult than estimating highway traffic conditions. To address these challenges, we propose a statistical modeling framework that leverages a large historical database and relies on the fact that traffic conditions tend to follow distinct patterns over the course of a week. This model is operational in North California, as part of the *Mobile Millennium* traffic estimation platform. The second goal of the article is to provide a global network-level analysis of traffic patterns using matrix factorization and clustering methods. These techniques allow us to characterize spatial traffic patterns in the network and to analyze traffic dynamics at a network scale. We identify traffic patterns that indicate intrinsic spatio-temporal characteristics over the entire network and give insight into the traffic dynamics of an entire city. By integrating our estimation technique with our analysis method, we achieve a general framework for extracting, processing and interpreting traffic information using GPS probe vehicle data.

1 Introduction and related work

Traffic congestion has a significant impact on economic activity throughout much of the world. Accurate, reliable traffic monitoring systems, leveraging the latest advances in technology and research are essential for active congestion control. They can also be used to study large scale traffic patterns and to understand specific travel behavior, network bottlenecks, to design long term infrastructure planning and to optimize mobility.

Until recently, traffic monitoring systems have relied exclusively on data feeds from dedicated sensing infrastructure (loop detectors and radars in particular). For highway networks covered by such infrastructure systems, it has become common practice to perform estimation of flow, density or speed at a very fine spatio-temporal scale [3], using traffic flow models developed in the last decades [32, 7, 41]. Probe vehicle data has also been successfully integrated into these models [45, 43, 24, 30]. For arterials, traffic monitoring is substantially more difficult: probe vehicle data is the only significant data source available today with the prospect of global coverage in the future. The lack of ubiquity and reliability, the variety of data types and specifications, and the randomness of its spatio-temporal coverage encourage the use of both historical and real-time data to provide accurate estimates of traffic conditions on large transportation networks. The *Mobile Millennium* project [26] receives probe vehicle data from a dozen of different sources. In Figure 1, we illustrate one of the data source of the *Mobile Millennium* project: it shows a snapshot of probe measurements from San Francisco taxis collected on an arbitrary day from midnight to 7:00am (small dots) as well as a snapshot of the probe locations at 7:00am (large dots). This figure illustrates both the breadth of coverage when aggregating data over long periods of time and the limited information available at a given point in time, limiting the direct estimation of the macroscopic state of traffic at a fine spatio-temporal scale. Note that filtering algorithms are designed to limit the bias of the different sources of data. For example, we filter the measurements during which the hired status of the taxis changes. Aside from less abundant sensing compared to existing highway traffic monitoring systems, the arterial network presents additional modeling and estimation challenges. The underlying flow physics is more complex because of traffic lights (often with unknown cycles), intersections, stop signs, parallel queues, and other phenomena.

We introduce a statistical approach for real-time arterial traffic estimation from probe vehicle data, leveraging massive amounts of historical data. Statistical approaches have been proposed that rely on either a single measurement per time interval or aggregated measurements per time interval [19, 10], neither of which is appropriate in our setting since probe data on arterials is available at random times and random locations. Some researchers have examined the processing of high-frequency probe data (one measurement approximately every 20 seconds or less) [43], which allows for reliable calculation of short distance speeds and travel times. In this article, we specifically address the processing of sparse probe data where this level of granularity is not available. Finally, other approaches based on regression [36], optimization [1], neural networks and pattern matching [8] have all been proposed. None of these approaches addresses the issue of processing sparse probe data on a dense arterial network.

Besides the ability of providing real time traffic estimation, the results produced by the model can be further analyzed to provide a large scale understanding of traffic dynamics both in time and in space. Most of previous research in traffic data analysis focus on temporal dynamics of individual links (either on arterial or highways) using data-driven approaches: in [39, 44, 35], Kalman filter and its extensions, originated from the theory framework of state space linear dynamic model, are used for modeling and tracking temporal variations of traffic flows; [46, 37] use neural networks to achieve short-term non-linear prediction of traffic flows based on historic observations; finally, [40] proposes to perform traffic prediction on individual links based on clustering of temporal patterns of traffic flows, while [11] adopts a time-series

Figure 1: San Francisco taxi measurement locations, observed at a rate of once per minute. Each small dot represents the measurement of the location of a taxi, received between midnight and 7:00am, on March 29th, 2010. The large dots represent the location of taxis visible in the system at 7:00am on that day.

analysis (Autoregressive Moving Average) on traffic flows in order to forecast traffic states. Very little progress has been made in analyzing the temporal dynamics of *global* traffic states of an *entire large-scale road network*. We call *global* traffic state, the aggregation of the congestion states of all the link of the network. Traffic states of neighboring individual roads are often highly correlated (both spatially and temporally) and the identification of specific traffic patterns or traffic configurations is very informative. They can be used to better understand global network-level traffic dynamics and serve as prior knowledge or constraints for the design of traffic estimation and prediction platforms. The analysis of traffic patterns is also useful for traffic management centers and public entities to plan infrastructure developments and to improve the performances of the available network using large-scale control strategies.

This article proposes an algorithm to identify spatial configurations of traffic states over the entire network and analyze large-scale traffic dynamics from traffic state estimates produced and collected over long periods of time. We define the *network-level traffic state* as the vector of traffic states for each link of the network at a given moment in time. It is represented in the form of multi-variate data, where its dimension is proportional to the amount of links in the transportation network. In large networks, this data structure quickly becomes too big to handle, limiting the analysis in the original high-dimensional space. In machine learning, this issue is commonly addressed using dimension reduction techniques (feature extraction) to simplify the representation of the data, remove redundancies and improve the efficiency of analysis techniques such as classification. Important applications of these algorithms include image processing and natural language processing [12, 9]. In this work, we propose to use a dimensionality reduction matrix factorization technique known as *Non-negative Matrix Factorization* (NMF) [6, 25] to obtain a low dimensional representation of network-level traffic states. Both the well-known *Principal Component Analysis* (PCA) method, and *Locality Preserving Projection* (LPP) technique are other examples of matrix factorization [28, 15]. However, in contrast to PCA or LPP, the NMF algorithm imposes strict non-negativity constraints on the decomposition result. This allows NMF to approximate the n-dimensional data vector by an *additive* combination of a set of learned bases. This property also leads to a part-based representation of the original data. The learned bases correspond to *latent components* of the original data so that the original data is approximated by a linear *positive* superposition of the latent components. The properties of the NMF have already been exploited for various applications. In text analysis, the learned bases are used to label different latent topics contained in text documents. In face image representation, the NMF bases indicate important localized components of the face, such as the eyes, the mouth or the cheeks. We expect that the distinctive characteristics of NMF will lead to a low-dimensional representation of network-level traffic states that exhibits global configurations of local traffic states and reflects intrinsic traffic patterns of network-level traffic

states.

The rest of this article is organized as follows. In Section 2 we present the real-time traffic estimation algorithm implemented in the *Mobile Millennium* system. It processes sparsely sampled probe vehicles sending their location at random places and random times and leverages historical data using a Bayesian update. In Section 3, we introduce the NMF algorithm, used in the remainder of the article to perform large scale analysis of the dynamics of traffic. In Section 4 and 5, we illustrate and provide a detailed analysis of typical spatial configuration patterns of network-level traffic states found by NMF projections. Section 6 further analyzes temporal dynamic patterns of the network-level traffic state, which describe evolutions of traffic states in the whole network. In Section 7, we conclude our work and discuss our future plans.

2 Large scale statistical model for arterial traffic estimation

We propose a parametric statistical model for large scale traffic estimation from sparsely sampled probe vehicles. The parameters of the model represent traffic patterns that are learned from massive amounts of probe vehicle travel times collected over long periods of time (section 2.1). The historic patterns are used as prior information in a Bayesian real-time estimation algorithm with streaming data (section 2.2). The statistical model is based on assumptions aimed at limiting the computational complexity of the algorithm while providing an adapted framework for arterial traffic estimation when little data is available in real-time but large quantities of historical data are collected over time.

1. The travel time on a link is a *random variable* (RV) and we assume that travel times on different links are independent RVs. In this article, we assume that travel times are normally distributed. Other distributions can also be used (e.g. Gamma, log-normal, and so on) without modifying the basic concepts of the algorithm. We will specify the equations that require modification under a non-normality assumption.
2. Any given moment in time belongs to exactly one *historic time period*, characterized by a day of the week, a start time and an end time. The set of historic time periods is denoted $\mathcal{T}$.
3. All travel time observations from a specific link l are independent and identically distributed within a given (historic) time period, $t \in \mathcal{T}$.
4. Probe vehicles send their location periodically (typically every minute). A trajectory reconstruction algorithm [27] provides the most likely path p of the vehicle between successive location reports. The path p is defined as a set of consecutive links, L_p, along with the fraction of the first and last link traversed and the total travel time associated with the entire path, y_p (time between successive location reports). The fraction of link l traversed for the pth observation on link l is denoted $w_{p,l}$. The time spent on link l is denoted x_p, l and have the constraint that $\sum_{l \in L_p} x_{p,l} = y_p$. The set of path observations for time period t is denoted $\mathbf{P}_t$. We assume that these path observations constitute the only data available to the model.
5. Each link of the network has a known minimum travel time b_l, found by considering the travel time that results from driving some percentage over the speed limit for the entire link. Note that the maximum travel time is harder to determine as travel times increase with congestion. A statistical analysis of available measurements may provide information about maximum travel times.

2.1 Learning historic traffic patterns

The historical model of arterial traffic estimates the parameters $Q_{l,t}$ of the travel time distribution on each link l for each historic time period t. The corresponding *probability density function* (PDF) of travel times is denoted $g_{l,t}(\cdot)$. In the case of Gaussian distributions, the parameters $Q_{l,t}$ are written $Q_{l,t} = (\mu_{l,t}, \sigma_{l,t})$, where $\mu_{l,t}$ and $\sigma_{l,t}$ represent the mean and the standard deviation of the travel times on link l for the period t.

For the pth path observation, the path travel time distribution is denoted $g_{L_p,t}(\cdot)$. Under assumption 1, the PDF of travel time on a path is computed as the convolution of the PDF of travel times of the links that make up the path.

The historic algorithm determines the values of $Q_{l,t}$ for each link and time period that are most consistent with the probe data received. This is achieved by maximizing the log-likelihood of the data given the parameters, which is written as

$$\arg\max_{\mathbf{Q}_t} \sum_{p \in \mathbf{P}_t} \ln(g_{L_p,t}(y_p)), \tag{1}$$

where $\mathbf{Q}_t$ is the set of $Q_{l,t}$ for all links l of the network. This optimization problem may be challenging due to the high number of variables (number of links times number of parameters per link travel time distribution), coupled through the PDF of path travel times $g_{L_p,t}$. In the case of Gaussian link travel times, solving (1) amounts to simultaneously estimating the mean and the variance of every link in the network and is not formulated as a convex problem. To face this difficulty, we decouple the optimization into two separate subproblems (*travel time allocation* [16, 20] and *parameter optimization*), each of which is easier to solve on its own, and then iterate between these subproblems until converging to an (locally) optimal solution.

If we knew how much time each probe vehicle drove on each link of its path (instead of just the total travel time), it would be easy to estimate the mean and standard deviation for each link in the network (sample mean and standard deviation of the link travel time observations). Since the sampling scheme only provides the total travel time on the path, we determine the most likely amount of time spent on each link (travel time allocation). Unfortunately, the most likely link travel times depend upon the link travel time parameters (μ and σ) that need to be estimated. This would appear to a chicken-and-egg problem, but there is a sound mathematical justification (hard EM) for iterating between these two steps. The link parameters are used to determine the most likely travel times and then the most likely travel times are used to update the parameters.

Travel Time Allocation: To solve the travel time allocation problem, we assume that estimates of the link parameters $Q_{l,t}$ are available (and fixed). We specify *lower bounds* b_l on the travel time allocated for each link l of the network to model the bounded speed of vehicles and ensure that a sensible solution is returned. For each observation $p \in \mathbf{P}_t$, we maximize the log-likelihood of the travel times $x_{l,p}$ spent on each link l of the path, with the constraint that they sum to the path travel time y_p. The optimization problem reads

$$\begin{aligned} \arg\max_{x} \quad & \sum_{l \in L_p} \log \mathcal{N}(x_{p,l};\ w_{p,l}\mu_{l,t}, w_{p,l}\sigma_{l,t}) \\ \text{s.t.} \quad & \sum_{l \in L_p} x_{p,l} = y_p \\ & x_{p,l} \geq w_{p,l} b_l, \forall l \in L_p, \end{aligned} \tag{2}$$

where the minimum travel time $w_{p,l}b_l$ is found using the minimum travel time b_l on link l, scaled by the fraction of link traveled $w_{p,l}$, as introduced in item 5. The notation $\mathcal{N}(x;\ \mu, \sigma)$ represents the PDF of a Gaussian variable with mean μ and standard deviation σ, evaluated at x:

$$\mathcal{N}(x;\ \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Problem (2) is a (small scale) quadratic program (QP) [4] which can be solved analytically (see algorithm 1 for details). Note that if a vehicle travels faster than the maximum speed, the allocation problem is infeasible. The vehicle is considered as an outlier and the observation is discarded from the set of observations. We call $\mathbf{X}_{l,t}$, the vector of allocated travel times for link l during time period t. Note that the travel times $x_{p,l}$ are scaled by the proportion of the link traveled, $w_{p,l}$, before being added to the set of allocated travel times $\mathbf{X}_{l,t}$.

Parameter Optimization: Given $\mathbf{X}_{l,t}$, the computation of the parameters $Q_{l,t}$ depends on the choice of the class of distribution chosen. In the case of Gaussian distributions, this computation is straightforward as $\mu_{l,t}$ and $\sigma_{l,t}$ respectively represent the sample mean and standard deviation of $\mathbf{X}_{l,t}$.

Full Historic Arterial Traffic Algorithm: After initializing the parameters $Q_{l,t}$ for each link of the network, the algorithm iterates between allocating the travel times for each path in $\mathbf{P}_t$ and optimizing the link parameters given the allocated travel times in $\mathbf{X}_{l,t}$. The convergence of the algorithm is checked by computing the log-likelihood after each iteration, which is guaranteed to increase at each iteration until convergence.

2.2 Bayesian Real-time Traffic Estimation

The parameters $Q_{l,t}$ learned by the historic model are used as prior information to estimate current traffic conditions via a *Bayesian update* (see [38] for more details on Bayesian statistics). In Bayesian statistics, parameters are considered as RVs and thus have a probability distribution. Here, we compute the probability distribution (known as *posterior distribution*) of the mean travel time (seen as a RV) given the allocated link travel times and a prior distribution on the mean travel time denoted f_0.

Let Δ_t represent the duration between successive real time estimates. We run the algorithm at time t_2 using the path data available for the *current time window* $[t_1, t_2]$, with $t_1 = t_2 - \Delta_t$. The duration Δ_t between successive updates depends upon the amount of data available in real-time and should remain inferior to the duration of the historical time intervals. If the data volume is large, the model can be run up to every 5 minutes. Running the model more frequently will likely not increase the performance and may lead to estimates that fluctuate too much due in particular to the periodic dynamics associated with the presence of traffic signals [22, 23].

For generic RVs X and Y with realization x and y, the notation $f(x|y)$ is read "probability that X has the realization x given that Y has the realization y" and denotes the conditional probability of RV X given the observation of the RV Y. Let y_{l,t_2} denote the set of travel times allocated to link l between t_1 and t_2. Using Bayes theorem, the posterior probability on the mean travel time $\hat{\mu}_{l,t_2}$ is proportional to the likelihood of the data times the prior:

$$f(\hat{\mu}_{l,t_2}|y_{l,t_2}, \sigma_{l,t}) \propto f(y_{l,t_2}|\hat{\mu}_{l,t_2}, \sigma_{l,t}) f_0(\hat{\mu}_{l,t_2}), \tag{3}$$

The symbol $\propto$ is read "is proportional to". The proportionality constant is chosen such that the integral of $\hat{\mu}_{l,t_2} \mapsto f(\hat{\mu}_{l,t_2}|y_{l,t_2}, \sigma_{l,t})$ on $\mathbb{R}$ is equal to one. At time t_2, the Bayesian update determines the value of mean travel times $\hat{\mu}_{l,t_2}$ that maximizes the posterior probability.

Assuming Gaussian link travel times, a natural choice for f_0 is a Gaussian distribution (conjugate prior [38]). Since f_0 represents prior information on the mean travel time, its mean is set to the historical mean $\mu_{l,t}$ and its standard deviation $\sigma_{0;\, l,t}$ is chosen to represent how much real-time condition can deviate from the historical values. Typically $\sigma_{0;\, l,t}$ is large to give more weight to real-time data as soon as they are in sufficient quantity. Because of the Gaussian prior, the allocated link travel times y_{l,t_2} and the mean travel time $\hat{\mu}_{l,t_2}$ are jointly Gaussian and we compute the parameters of the posterior (Gaussian) distribution of $\hat{\mu}_{l,t}$. In particular, we update the mean link travel times as the mean of the posterior distribution:

Algorithm 1 Travel time allocation algorithm. The core of the algorithm is contained in lines 11-15, which computes the total expected path variance (V) and the difference between expected and actual travel times (Z). With these two quantities, each link is allocated the expected link travel time adjusted by some proportion of Z, where this proportion is computed using the link variance divided by the total path variance. This procedure can lead to some links being allocated a travel time below the minimum for that link. The set $\mathbf{J}$ is introduced to track the links with initial allocated travel times below the lower bound and the main procedure is repeated by setting the travel times for these links to the lower bound and optimizing with respect to the remaining links. Note that the travel times are scaled by the proportion of the link traveled (line 19) before being added to the set of allocated travel times $\mathbf{X}_{l,t}$.

Require: $t \in \mathcal{T}$ is fixed to some particular time period.

1: **for** $l \in \mathcal{L}$ **do**
2: $\mathbf{X}_{l,t} = \emptyset$ {Initialize allocated travel time sets to be empty.}
3: **end for**
4: **for** $p \in \mathbf{P}_t$ **do** {For all probe path observations.}
5: **if** $\sum_{l \in L_p} w_{p,l} b_l > y_p$ **then**
6: Travel time allocation infeasible for this path. This means that the observation represented travel that is considered faster than realistically possible, so the observation is considered an outlier. Remove p from $\mathbf{P}_t$.
7: **else**
8: $\mathbf{J} = \emptyset$ {$\mathbf{J}$ contains all links for which the travel time allocation is fixed to be equal to the lower bound.}
9: **repeat**
10: $x_{p,l} = w_{p,l} b_l, \forall l \in \mathbf{J}$ {For all links that had an infeasible allocation in the previous pass through this loop, set the allocation to the lower bound.}
11: $V = \sum_{l \in L_p \setminus \mathbf{J}} w_{p,l} \sigma^2_{l,t}$ {Calculate the path variance for the links not fixed to the lower bound.}
12: $Z = y_p - \sum_{l \in \mathbf{J}} w_{p,l} b_l - \sum_{l \in L_p \setminus \mathbf{J}} w_{p,l} \mu_{l,t}$ {Calculate the difference between expected and actual travel time for the links not fixed to the lower bound.}
13: **for** $l \in L_p$ **do** {Allocate excess travel time in proportion of link variance to path variance.}
14: $x_{p,l} = w_{p,l} \mu_{l,t} + \frac{w_{p,l} \sigma^2_{l,t}}{V} Z$
15: **end for**
16: $\mathbf{J} = \mathbf{J} \cup \{l \in L_p : x_{p,l} < w_{p,l} b_l\}$ {Find all links violating the lower bound.}
17: **until** $x_{p,l} >= w_{p,l} b_l, \forall l \in L_p$
18: **for** $l \in L_p$ **do**
19: $\mathbf{X}_{l,t} = \mathbf{X}_{l,t} \cup \left(\frac{x_{p,l}}{w_{p,l}} \right)$ {Add the allocated travel time to $\mathbf{X}_{l,t}$.}
20: **end for**
21: **end if**
22: **end for**
23: **return** $\mathbf{X}_{l,t}, \forall l \in \mathcal{L}$

$$\hat{\mu}_{l,t} = \frac{\sigma^2_{0;\,l,t}}{\frac{\sigma^2_{l,t}}{N_{l,t_2}} + \sigma^2_{0;\,l,t}} \overline{x} + \frac{\sigma^2_{l,t}}{\frac{\sigma^2_{l,t}}{N_{l,t_2}} + \sigma^2_{0;\,l,t}} \mu_{l,t}$$

where N_{l,t_2} is the number of travel times allocated to link l during the current time interval (t_1, t_2) and $\overline{x}$ is the sample mean of the allocated travel times y_{l,t_2}.

To summarize, the real-time estimation algorithm performs the travel time allocation on each probe observation and then uses the allocated travel times and the historical traffic parameters to perform a Bayesian update of the link parameters.

The precise analysis of the performance of this model is out of the scope of this article. We refer the reader to the following references assessing the results of the *Mobile Millennium* project for more details [2, 17].

3 Non-negative matrix factorization (NMF)

In this section, we present *Non-negative Matrix Factorization* (NMF), which is used for approximating network-level traffic states as positive sums of a limited number of global traffic configurations. NMF [31, 6, 33, 25, 9] is a particular type of matrix factorization, in the same domain as the well-known *Principal Component Analysis* (PCA) method and *Locality Preserving Projection* (LPP). In all cases, given a set of multivariate n-dimensional data vectors placed in m columns of a $n \times m$ matrix X, matrix factorization decomposes the matrix into a product of a $n \times s$ loading matrix M and a $s \times m$ score matrix V, where s represents the dimensionality of the subspace to which we project the original data. Through this matrix decomposition, each n-dimensional data vector is approximated by a linear combination of the s columns of M, weighted by the components in the corresponding column of V. We can regard all s column vectors in loading matrix M as a group of projection bases that are learned optimally to represent the original data. The variable s is typically chosen to be significantly smaller than both n and m so that the obtained score matrix V forms a low-dimensional subspace projection of the network-level traffic states, on which we can perform further data analysis. The specificity of NMF is the enforced positivity of both the weights in V, and of the columns of M forming the NMF decomposition basis. This non-negativity therefore provides an approximation of the n-dimensional data vector by an *additive* combination of a set of learned bases. Furthermore, the NMF components forming the basis tend to be sparse, which leads leads to a part-based representation of the original data.

A network-level traffic state is a vector of size equal to the number of links in the network, where the i^{th} entry corresponds to the traffic state on the i^{th} link of the network. Arterial networks are typically dense (numerous links and intersections) and the number of links in any decent size network is often over a thousand links. Assuming that k samples of n-dimensional network-level traffic states are stored as an $n \times k$ matrix X, NMF factorizes X as a product of a non-negative $n \times s$ matrix M and a non-negative $s \times k$ matrix V which minimizes the Frobenius norm of the reconstruction error between X and its factorized approximation MV. We recall that the Frobenius norm of a matrix $A \in \mathbb{R}^{n \times m}$ with entry on column i and line j denoted $A_{i,j}$ is defined as

$$||A||_F = \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{m} |a_{i,j}|^2}$$

and is equal to the sum of the singular values of A. The matrix factorization problem reads:

$$\arg \min_{(\mathbf{M},\mathbf{V})} \|\mathbf{X} - \mathbf{MV}\|_F \text{ s. t. } \quad M \geq 0,\ V \geq 0, \tag{4}$$

where the inequalities $M \geq 0,\ V \geq 0$ represent the non-negativity constraints (each element of the matrices are non-negative). Training of NMF is implemented using multiplicative updates [31],fixing either M or V and updating the left following the KKT condition. The NMF cost function shown in equation (4) is not convex.However,fixing either M or V leads to a convex subproblem to solve. Multiplicative updates and other gradient based optimization procedure can not guarantee the global optimum of the NMF solution. Nevertheless, in data mining, local minimum is still enough to be useful. Given fixed M or V, the NMF objective is a convex optimization issue. The NMF projects the high-dimensional network-level traffic states on a s-dimensional subspace, which is spanned by the columns of M. According to equation (4), the column space of V corresponds to coordinates of network-level traffic states with respect to the learned set of bases in M. The column space of V forms a low-dimensional representation of the network-level traffic states. As mentioned in the introduction, each network-level traffic state $X_j \in \mathbb{R}^n$ is approximated by an additive linear superposition of the column space of M due to the non-negative constraint. The approximation of X_j is written

$$X_j \approx \sum\nolimits_{i=1}^{k} M_i V_{i,j}, \tag{5}$$

where M_i denotes the i^{th} column of M and $V_{i,j}$ is the element at the i^{th} column and j^{th} row of V. It is important to interpret what the matrices M and V represent in terms of traffic analysis. The column space of M represents typical elements of the spatial configuration patterns with respect to the network-level traffic states. Based on the columns of M, we represent complex spatial arrangements of local traffic states over the entire network. As for V, equation (4) indicates that each element $V_{i,j}$ represents to which degree the j^{th} network-level traffic state observation is associated with the i^{th} expanding basis in matrix M (i^{th} spatial configuration). For example, if the spatial configuration formed by the i^{th} column of M is the best representation of the j^{th} network-level traffic state, then $V_{i,j}$ will take the largest value in the j^{th} row of V [6]. As a result, the derived low-dimensional representation formed by the columns space of V are intuitively consistent with information about spatial distribution patterns of local traffic states. By contrast, the PCA and LPP based projections only aim at best reconstruction of traffic observations with either maximizing data variances or preserving neighboring structures. The projection results of PCA and LPP are thus less likely to be associated with interpretable latent traffic configuration patterns than the NMF. Therefore, we choose NMF to analyze the network-level traffic states in our case.

In this article, the traffic states used for the clustering analysis are *fluidity indices*. A fluidity index is a value in $[0, 1]$ computed as the ratio between the free flow and the estimated travel times. They are provided by the estimation algorithm described in Section 2 and operational in the *Mobile Millennium* [26] traffic platform, which receives data from a dozen of feeds totalling several millions of data points per day for Northern California traffic. The *Mobile Millennium* platform has been operational since November 2008 and has been storing historical data since then, providing a rich database of historical traffic dynamics in the Bay Area. In real-time, the model estimates travel times and fluidity indices from the streaming data and leverages the historical data using the Bayesian update presented in Section 2.2. The estimates are updated on each link of the network every five minutes. We focus our study on a network consisting of 2626 links for a duration of 184 days, from 00:00 May 1st 2010 to 23:55 October 31st 2010, totaling 52292 estimates per link ($12 \times 24 \times 184$). We store the fluidity index of each link at each time sampling step in a matrix X containing 2626 rows and 52292 columns. Our clustering results includes two parts, firstly we perform clustering on network-level traffic states, in order to find some typical spatial configurations of network-level traffic states, as described in Section 4. Secondly, we perform clustering on temporal trajectories of network-level traffic states, from which we can study traffic dynamics, shown in Section 6.

4 Congestion patterns: spatial configurations of global traffic states

An important outcome of dimensionality reduction is to identify typical spatial congestion patterns (i.e. spatial configurations of congestion). While doing this on the original 2626 dimensional data would be rather sloppy and computer intensive, it is much more feasible in the low dimensional space obtained by NMF.

NMF has one essential parameter: the number s of components over which decomposition is done. The parameter s also corresponds to the dimension of the target subspace where we perform clustering. The choice of s is empirical (s is called a *meta parameter*) and is done by analyzing results obtained for increasing values of s from 3 to 30. Our analysis focuses on the reconstruction error (value of the objective function (4) at optimum) and the clustering results. The reconstruction error continually decreases as the dimension s increases. This result is expected as the optimization problem (4) is performed on a larger set and thus the factorization models with higher complexity always leads to better fitting to the original data. Our clustering of global traffic states consists of clustering the traffic data projected in the s-dimension subspace using a *k-means* algorithm [34, 29]. The k-means algorithm is a widely used unsupervised clustering algorithm. It partitions observations into k clusters in which each observation belongs to the cluster with the nearest mean. We represent the clusters obtained in the s-dimensional space in three dimensions, limiting the number of NMF components to three but keeping the clustering results obtained in the s-dimensional space. We notice that values of s inferior to eleven lead to clustering results which seem visually inadequate: the 3D representation of the clusters shows important overlap between the clusters. The clusters become separated for values of s greater than fifteen. Increasing s over 15 does not seem to bring any improvement in the clustering results, while it significantly increases the NMF computation and memory usage costs. Therefore, we set the number of NMF components to $s = 15$ for all subsequent analysis presented in this article. This value achieves a balance between the descriptive power of NMF projection and the computational efficiency.

In clustering analysis, we also need to choose the number of clusters in k-means, denoted by k. The challenge is different from that for the choice of s: the choice of k does not influence the computational costs significantly but changes the interpretability of the results. The number of clusters represents the number of "global congestion patterns" that may arise. Too low values of k may not represent the different congestion patterns whereas too high values of k may decrease the possibilities of interpretation by separating similar congestion states into different clusters. After analyzing the results obtained for increasing values of k, it seems that the most insightful clustering is obtained with $k = 5$ clusters. The average fluidity index value (obtained by averaging index values on all links) are shown for each of the five clusters in the table at the top of figure 2. It appears that two clusters (cyan squares and green stars) correspond to different types of "mostly fluid" states, whereas the remaining three clusters (blue circles, yellow diamonds and red stars) represent "congested states". We study the physical significance of each cluster by constructing histograms of the fluidity index values, counting occurrence frequencies of fluidity index values in each cluster. We find that fluidity index values in the *night and early morning Free-Flow* (NFF) and *Evening Free-Flow* (EFF) cluster are higher as a whole than those in the clusters corresponding to occurrences of congestion (*Morning Increasing Congestion*, MIC, *Mid-Day Congestion*, MDC and *Afternoon Decreasing Congestion* ADC clusters).

Figure 2 shows that the significance of the distributional patterns with respect to evaluating global traffic states is generally consistent with that of average fluidity index values, which implies that the average fluidity index value could also be used as an easy-to-use and efficient indicator of global traffic states in our case.

Marker symbol	Average fluidity	Cluster name
Green stars	0.7757	Night + early morning Free-Flow (NFF)
Blue circles	0.7185	Morning Increasing Congestion (MIC)
Red stars	0.6393	Mid-Day Congestion (MDC)
Yellow diamonds	0.6730	Afternoon Decreasing Congestion (ADC)
Cyan squares	0.7420	Evening Free-Flow (EFF)

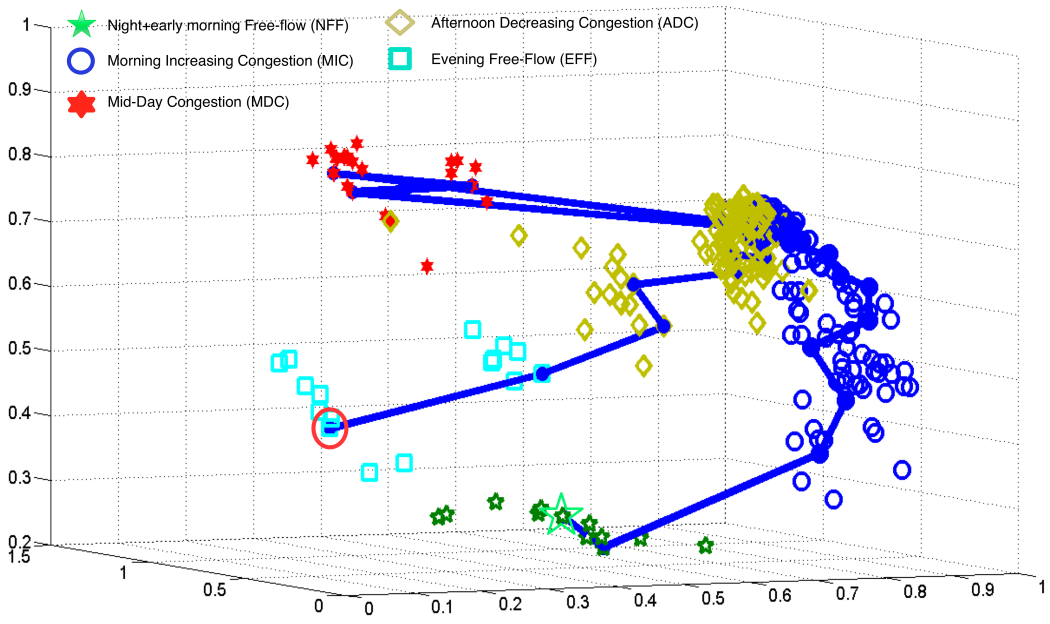
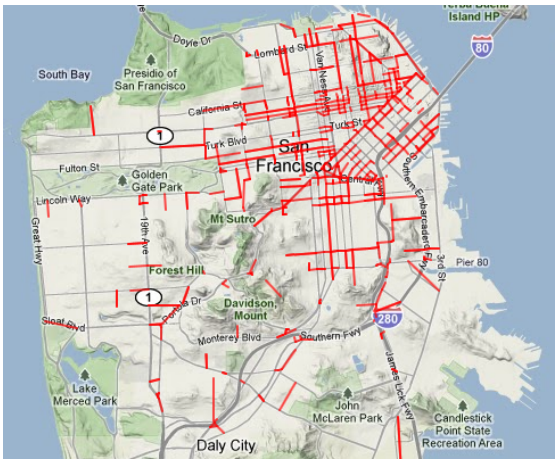
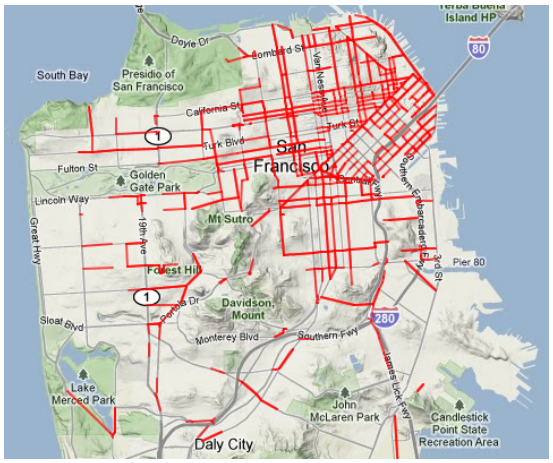


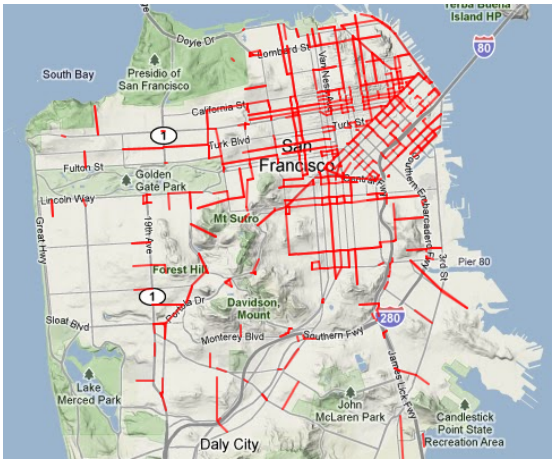
Figure 2: The clustering shows an organization of global congestion states per time of the day. The table shows the average fluidity values of each of the global state clusters. The figure shows the temporal evolution of global congestion states, projected in the 3D-NMF space using different colors and symbols to represent the five different clusters. The first and the last network estimates of the day are represented with a large star and a large circle respectively.



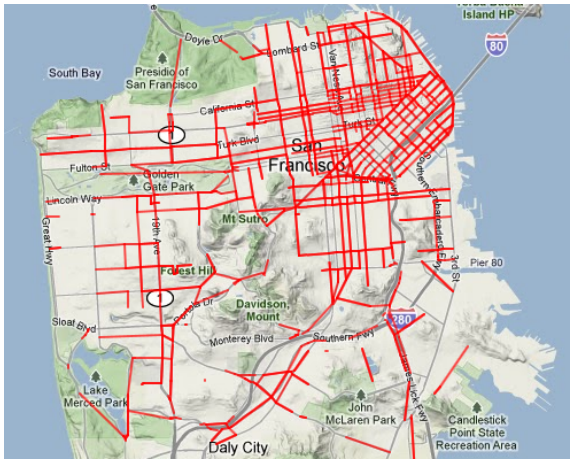
(a) Night Free-Flow (NFF)



(b) Morning Increasing Congestion (MIC)



(c) Evening Free-Flow cluster (EFF)



(d) Afternoon Decreasing Congestion (ADC)

(e) Mid Day Congestion (MDC)

Figure 3: Typical spatial configurations of traffic states for each of the five clusters. On each figure, we display the links with fluid index values less than 0.7 (congested links). Most of the congestion occurs within the region highlighted by the dashed circle, which is the downtown region of San Francisco. The NFF and EFF clusters have a smaller number of links highlighted than the MIC, ADC and MDC clusters indicating the difference in congestion levels.

As done in the primary analysis for the choice of s and for visualization purposes, we illustrate spatial layouts of the global traffic state distribution in 3D-NMF space (obtained by requesting 3 components only instead of 15), but we apply the k-means clustering algorithm in the larger 15-D NMF space. The physical interpretation of the five clusters is clear in Figure 2 in which we show all states projected in 3D-NMF, together with a typical temporal evolution trajectory of a single day. The whole trajectory is indicated by the blue line in Figure 2. The green star and red circle are the starting point and ending point of the trajectory,corresponding to traffic observations at 00:00 and 23:55 respectively. The temporal arrangements of the network-level traffic states along the trajectory underline the temporal interpretation of the five clusters: the green-star cluster corresponds to night and early morning free-flowing, from which typical day evolution goes into morning intermediate states (before 10:00) corresponding to the blue-circle cluster; mid-day congestion (red-star cluster) generally occurs between 10:05 and 15:00, and represents a clearly different congestion state in 3D-NMF space, with a sudden jump of traffic states from the blue-circle cluster to the red-star one, and sudden jump back into the afternoon intermediate state (yellow-diamond cluster) around 15:00. The traffic settles to a specific evening near-free-flow state from 18:00 to 23:55 (cyan-square cluster). Interestingly, both the projection of the global congestion states in 3D-NMF space and the clustering results in 15D-NMF show a clear distinction between morning and afternoon intermediate congestion states, and also between late evening and night/early-morning near-free-flow states.

In Figure 3, we show traffic patterns corresponding to spatial configurations of congestion for centers of each of the five identified clusters. Each cluster center is derived by averaging all elements of the corresponding cluster, so as to indicate a representative spatial configuration of traffic states of each cluster. In this figure, we display the links with fluidity index values less than 0.7 (congested links) on the Google Map screenshots. Generally, most of congestion occurs within the regions highlighted by the dashed circle in figure 3(e). This region corresponds to the downtown region of San Francisco. Compared with the downtown region, the western and southern region of San Francisco are less likely to suffer from congestion (left and bottom part in the San Francisco road network). This analysis is very useful for traffic management centers and public entities to understand the most important bottlenecks that cause heavy traffic conditions. Moreover, our results show that some of the major bottlenecks remain constant throughout the day whereas others evolve with the different traffic patterns of the day. This dynamical analysis can lead to specific management strategies to address this recurring congestion. As a matter of fact, in [13], we constructed a regression model to predict the global traffic dynamics based on the analysis results of the spatial congestion patterns. This work indicates the promising potentials of spatial congestion patterns in forecasting congestion and improving traffic management.

5 Spatial decomposition of the road network

Another motivation for using NMF in dimensionality reduction is its property to approximate original data by an additive linear combination of a limited set of "components" (a.k.a. NMF "basis"). Due to the non-negative constraints, spatial arrangements of the components are usually sparse, which means that values in most regions of each basis are (close to) zero except several localized regions. These localized regions with large values correspond to typical patterns or representative components of the original signals (the global congestion states), and typically correspond to independent "parts" of the data. Therefore, NMF is often used to extract part-based representation or latent semantic topics from the data in image processing or text classification. For example, when NMF is applied to image datasets, it automatically extracts some part-based representation of the type of objects present in the images [31, 25, 9]. We study this "part-based" representation of global congestion states to analyze the physical

significance of NMF components obtained on traffic data.

For arterial traffic, the localized regions with distinctively large values in each NMF basis correspond to a group of links with highly correlated traffic states. In this section, we construct the localized components by selecting the links which represent the top 20% largest values in each basis and indicate their spatial locations using red legends in the road network. Figure 4 shows several typical spatial arrangements of localized components, out of the fifteen arrangements learned during the NMF training. We notice that a component corresponds to streets in a localized West region (Figure 4(a)), and another to streets in the central region (Figure 4(b)), which could indicate that the traffic within each of these regions is highly correlated with each other whereas the traffic between distinct regions exhibits relatively independent behaviors. Such a characterization of independent regions of traffic dynamics is important to significantly reduce the computational costs of a large variety of estimation models, in particular estimation models based on graphical models [10, 18]. We could leverage this characterization in approximate inference algorithms to reduce the computational costs while maintaining an accurate representation of traffic dynamics and limiting the estimation error [5].

Other NMF components highlight correlations of traffic in parallel directions: in Figure 4(c), a majority of the links of the NMF component are horizontally-oriented, wheras in Figure 4(d) a majority of the links are vertically-oriented. As we highlight in Figure 4(c) and Figure 4(d),the links concentrated within the downtown tend to be more consistent with the orientational patterns. These links with similar orientations are likely to have correlated traffic dynamical behaviors, whereas traffic flows with orthogonal orientations have a less important impact on each other. These correlations properties can be used to learn the structure of the graphical model representing conditional independences between traffic states on the network (both spatially and temporarily).

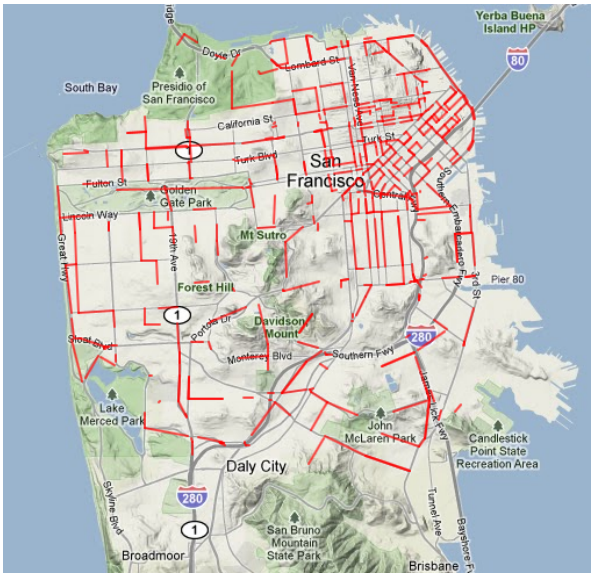
According to the physical representation of the NMF components, it seems that different NMF bases focus on different localized connected regions of the network. This could imply that NMF detects both strong correlation of traffic dynamics within each localized region and relative independence between these regions. However, this connectivity and localization of the components could be improved. Standard NMF does not guarantee connected nor localized components and the above promising results motivate us to investigate this physical representation of spatial configuration of traffic states further. A possible approach is to modify the NMF algorithm in order to favor *localized* sparsity, which should help to unveil more distinct part-based network decomposition.

6 Temporal analysis of global traffic states

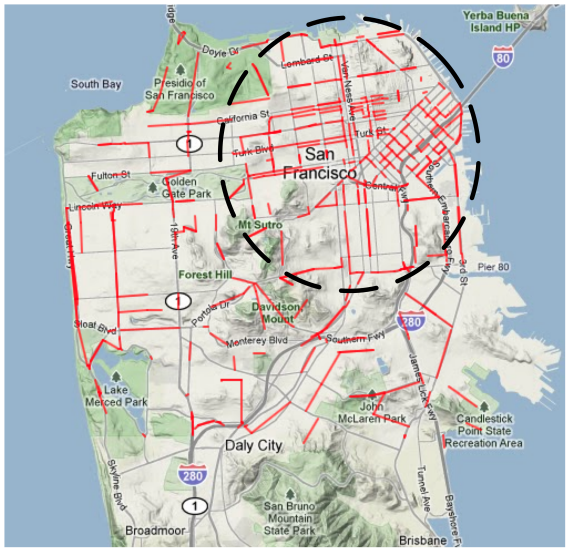
In this section, we analyze the daily dynamics of network-level congestion states projected in the NMF space. This analysis is important to understand how congestion forms and dissipates throughout the day. For each day in the studied period, we represent the trajectory of the network-level traffic states in the NMF space as the projection of the temporal sequence of the network-level traffic states from the beginning to the end of the day. The projections are linked together to form a solid curve representing the trajectory and we notice that trajectories are nearly closed in the NMF space. Note that for visualization purposes, the projection is done on the 3D-NMF space. Figure 5 (top) shows a typical day trajectory with successive temporal intervals along the trajectory plotted using different colors, to give an idea of the dynamics along the curve.

It is noteworthy that over the 184 days of reconstructed traffic data, there are only, in 3D-NMF projection, exactly seven different typical trajectories, as shown in figure 5 (center). Furthermore, our analysis shows that each one of these seven typical trajectories corresponds to

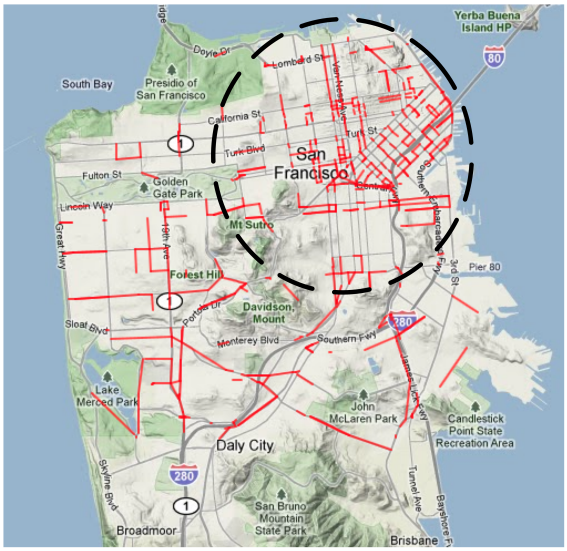
(a) "West Part" NMF component



(b) "Central" NMF component



(c) "East-West transit" NMF component



(d) "North-South transit" NMF component

Figure 4: Examples of interesting NMF components, either highlighting localized behavior (a and b), or flow-direction correlations (c and d).

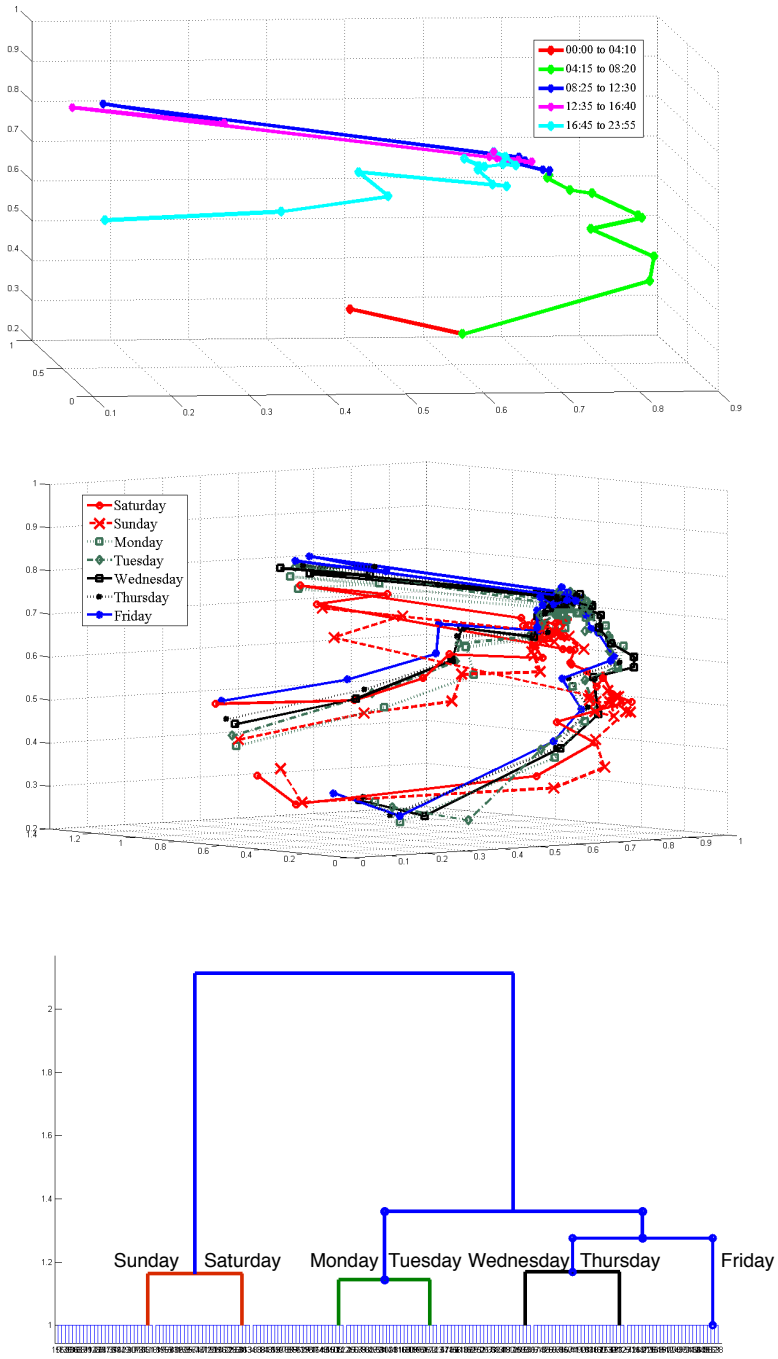


Figure 5: Daily trajectories of network fluidity indices projected in 3D-NMF space exhibit seven different typical trajectories, representing the days of the week. **Top**: Example of a daily trajectory with coloring representing the five different times of the day. **Center**: The seven different trajectories, representing a typical daily dynamic for each day of the week. **Bottom**: Dendrogram representing the hierarchical clustering analysis of the daily trajectories.

a particular day of the week and are thus called *day trajectory patterns*. Note that individual day trajectories for same day-of-week, although superposed in 3D-NMF, are slightly different one from another in 15D-NMF space, in which we perform clustering. Also, there are only two weekday holidays within the analyzed period, and they exhibit only small deviation from the ordinary same day-of-week. This may be a consequence of the estimation algorithms which does not use a holiday-specific model to process the historic data and fuse it with the real-time data.

Differences between the different day trajectory patterns concentrate within the time interval corresponding to transitions between congestion states, in particular between the morning increasing congestion and the mid-day congestion and between the mid-day congestion and the evening decreasing congestion. Characterizing these specific time intervals that represent the differences in daily dynamics allows us to identify and/or predict different evolution patterns of traffic states and to develop mid-term or long-term traffic forecast [14, 13].

In this data set, one complete evolution trajectory contains 288 sampling steps (estimations are performed every five minutes), which is represented by a 2626×288 matrix (the network has 2626 links). As for the previous sections, our analysis is done in 15-D NMF space (3-D space is only used for visualization purposes). Each trajectory is represented by a sequence of 288 network-level traffic state projected on the 15-D NMF space and denoted $\{h_1, h_2, \cdots, h_{288}\}$, where $h_i \in \mathbb{R}^{15}$. To measure similarity between trajectories $\{h_1^a, h_2^a, \ldots, h_{288}^a\}$ and $\{h_1^b, h_2^b, \ldots, h_{288}^b\}$, representing days a and b respectively, we calculate *cosine distances* between the NMF projections at corresponding times of the day and sum the cosine distances over the different estimation times $k = 1 \ldots 288$:

$$D = \sum\nolimits_{k=1}^{288} \text{cosdis}(h_k^a, h_k^b), \tag{6}$$

$$\text{where cosdis}(h_k^a, h_k^b) = 1 - \frac{h_k^a \cdot h_k^b}{\| h_k^a \| \, \| h_k^b \|}. \tag{7}$$

The function *cosdis* is the cosine distance between two vectors and is defined in (7). It evaluates the cosine value of the angle between the two data vectors h_k^a and h_k^a in the 15-D NMF projection space. Larger cosine distance values indicate more important differences between the two vectors. Due to the mathematical definition of the cosine function, the derived cosine distance is normalized into the range [0,1]. Based on the defined distance measure between sequences, we can perform hierarchical clustering of daily traffic observation sequences in 15D-NMF space [29, 42]. The successive similarity-based groupings are shown on the dendrogram in Figure 5 (bottom) following the same color legends as in the middle figure. In the dendrogram, daily sequences of network-level traffic states are grouped gradually into clusters in the form of U-shaped trees. The height of each U-shaped tree (vertical axis) represents the distance between the sets of daily sequences being connected. Leaf nodes along the horizontal axis correspond to all daily sequences of network-level traffic states. We notice that at the bottom level of the hierarchical tree, daily sequences are first aggregated with respect to each day of the week. It underlines the intuition that each day of the week has a particular temporal dynamic pattern in terms of network-level traffic states. By increasing thresholds of distance settings, we trace back along the U-shaped trees towards its root. The seven days of the week are further clustered into four different groups indicating the days that tend to follow similar dynamic patterns. Weekend (Saturday and Sunday) are clustered together. As for the week days, Monday and Tuesday, representing the beginning of a week, appear to have a different temporal dynamic pattern from Wednesday and Thursday (middle of the week). Traffic dynamics on Fridays also tend to deviate slightly from that of the other days and is assigned to a separate group. As the distance threshold increases, Friday is added to the Wednesday and Thursday cluster. Therefore, we can say that there are generally three kinds of temporal dynamic patterns of network-level traffic states in the data, corresponding to the beginning of the week (Monday and Tuesday), the end

of the week (Wednesday, Thursday and Friday) and weekends (Saturday and Sunday). If we increase the threshold even more, the two clusters of week days merge leading to two clusters representing the week-end days on one side and the week days on the other side. The distance thresholds need to be increased significantly more for these two clusters to merge, which indicates the importance in the differences in daily dynamics between week days and weekends. It is expected for Monday and Friday to have different dynamics (coming back or leaving for the week-end). However, it is slightly surprising that Monday and Tuesday are clustered together while Wednesday and Thursday (and then Friday) form another cluster. The data seems to indicate a beginning of the week vs. end of the week clustering, with Friday being the most different of the other days.

7 Conclusion and discussion

In this article, we have proposed and presented: (1) a probabilistic modeling framework for efficient estimation of arterial traffic conditions from sparse probe data; (2) a novel traffic data mining approach to analyze large-scale traffic patterns and dynamics.

The proposed estimation method leverages massive amounts of historical data to learn statistical distributions of travel times and fuses them with streaming data to produce real-time estimates of traffic conditions using a Bayesian update. The Bayesian framework allows us to properly weight the relative importance of the real time and the historical data (depending on the amount of data available in real time) to produce robust estimates, even when little data is available in real-time. This model is operational in the *Mobile Millennium* system and has been producing travel time and fluidity indices since March 2010 [2].

The output of this estimation model is used as a first real-world platform for a new traffic data mining method using Non-negative Matrix Factorization to allow large-scale analysis of spatial and temporal traffic patterns. The principle is to perform dimensionality reduction, which allows for clustering of spatial congestion patterns, and easy analysis/categorization of temporal daily dynamics. Furthermore, the part-based decomposition feature of Non-negative Matrix Factorization automatically unveils areas of the road network with strong correlations.

Current and future research focus on: (1) integrating traffic flow theory and statistical models to have a more accurate modeling of traffic dynamics, both at the link [22, 23] and at the network [21] level in order to improve the estimation capabilities of the system; (2) modifications of Non-negative Matrix Factorization sparsity constraint to favor geographically-localized components; (3) taking advantage of low-dimensional Non-negative Matrix Factorization representation for performing long-term traffic prediction [13].

Acknowledgements

The authors wish to thank Timothy Hunter from UC Berkeley for providing filtered probe trajectories from the raw measurements of the probe vehicles. We thank the *California Center for Innovative Transportation* (CCIT) staff for their contributions to develop, build, and deploy the system infrastructure of *Mobile Millennium* on which this article relies. This research was supported by the Federal and California DOTs, Nokia, the Center for Information Technology Research in the Interest of Society (CITRIS), the French National Research funding Agency ANR (part of this work was supported by grant ANR-08-SYSC-017 for the "TRAVESTI" project), and French Department of Sustainable Development and Transports MEEDDM. This collaboration between CAOR of Mines ParisTech and Berkeley was initiated partly thanks to funding MEEDDM-09-SDI-003 granted by French authorities within the CalFrance franco-californian cooperation framework.

References

[1] S.J. Agbolosu-Amison, B. Park, and I. Yun. Comparative evaluation of heuristic optimization methods in urban arterial network optimization. In *12th Intelligent Transportation Systems Conference (ITSC '09)*, 2009.

[2] A. Bayen, J. Butler, and A. Patire et al. Mobile Millennium final report. Technical report, University of California, Berkeley, CCIT Research Report UCB-ITS-CWP-2011-6, To appear in 2011.

[3] P. Bickel, C. Chen, J. Kwon, J. Rice, E. Van Zwet, and P. Varaiya. Measuring traffic. *Statistical Science*, 22(4):581–597, 2007.

[4] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[5] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proc. UAI*, volume 98, 1998.

[6] D. Cai, X. He, X. Wu, and J. Han. Non-negative matrix factorization on manifold. In *The proceedings of ICDM 2008*, 2008.

[7] C. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research B*, 28(4):269–287, 1994.

[8] C. de Fabritiis, R. Ragona, and G. Valenti. Traffic estimation and prediction based on real time floating car data. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 197–203, 2008.

[9] T. Feng, S.Z. Li, H.Y. Shum, and H.J. Zhang. Local nonnegative matrix factorization as a visual representation. In *Proceedings of the 2nd International Conference On Development and Learning*, 2002.

[10] C. Furtlehner, J. Lasgouttes, and A. de la Fortelle. A belief propagation approach to traffic prediction using probe vehicles. In *10th Intelligent Transportation Systems Conference (ITSC '07)*, pages 1022–1027, 2007.

[11] B. Ghosh, B. Basu, and M. O'Mahony. Multivariate short-term traffic flow forecasting using time-series analysis. *IEEE Transaction on Intelligence Transportation Systems*, 10(2):246–254, 2009.

[12] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[13] Y. Han and F. Moutarde. Analysis of network-level traffic states using locality preservative non-negative matrix factorization. In *14th IEEE Intelligent Transport Systems Conference (ITSC'2011)*, 2011.

[14] Y. Han and F. Moutarde. Clustering and modeling of network-level traffic states based on locality preservative non-negative matrix factorization. In *8th Intelligent Transport Systems (ITS) European Congress*, 2011.

[15] X. He and P. Niyogi. Locality preserving projections. In *Proceedings of 17th Neural Information Processing Systems*, 2003.

[16] B. Hellinga, P. Izadpanah, H. Takada, and L. Fu. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C*, 16(6):768 – 782, 2008.

[17] R. Herring. *Real-Time Traffic Modeling and Estimation with Streaming Probe Data using Machine Learning*. PhD thesis, UC Berkeley, Departement of Industrial Engineering and Operations Research, 2010.

[18] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen. Estimating arterial traffic conditions using sparse probe data. In *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*, Madeira, Portugal, September 2010.

[19] R. Herring, A. Hofleitner, S. Amin, T. Abou Nasr, A. Abdel Khalek, P. Abbeel, and A. Bayen. Using mobile phones to forecast arterial traffic through statistical learning. In *Proceedings of the 89th Annual Meeting of the Transportation Research Board*, Washington D.C., 2010.

[20] A. Hofleitner and A. Bayen. Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model. *IEEE Intelligent Transportation System Conference (IEEE ITSC '11)*, 2011.

[21] A. Hofleitner, R. Herring, and A. Bayen. Arterial travel time forecast with streaming data: a hybrid flow model - machine learning approach. *submitted, Transportation Research Part B*, 2011.

[22] A. Hofleitner, R. Herring, and A. Bayen. A hydrodynamic theory based statistical model of arterial traffic. *Technical Report UC Berkeley, UCB-ITS-CWP-2011-2*, January 2011.

[23] A. Hofleitner, R. Herring, and A. Bayen. Probability distributions of travel times on arterial networks: a traffic flow and horizontal queuing theory approach. *91st Transportation Research Board Annual Meeting*, January 2012.

[24] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005.

[25] P.O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, vol.5:1457–1469, 2004.

[26] The *Mobile Millennium* Project. http://traffic.berkeley.edu.

[27] T. Hunter, R. Herring, A. Hofleitner, A. Bayen, and P. Abbeel. Trajectory reconstruction of noisy GPS probe vehicles in arterial traffic. *In preparation for IEEE Transactions on Intelligent Transport Systems.*

[28] I.T. Jolliffe. *Principal component analysis*. Springer, 2002.

[29] T. Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, and A.Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans Pattern Analysis and Machine Intelligence*, vol.24:881–892, 2002.

[30] A. Krause, E. Horvitz, A. Kansal, and F. Zhao. Toward community sensing. In *Proceedings of ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, St. Louis, MO, April 2008.

[31] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS 2000, pp.556-562*, 2000.

[32] M. Lighthill and G. Whitham. On kinematic waves. II. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 229(1178):317–345, May 1955.

[33] C.J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, vol.19,no.10:2756–2779, 2007.

[34] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[35] W. Min, L. Wynter, and Y. Amemiya. Road traffic prediction with spatio-temporal correlations. Technical report, IBM Watson Research Center, 2007.

[36] X. Min, J. Hu, Q. Chen, T. Zhang, and Y. Zhang. Short-term traffic flow forecasting of urban network based on dynamic STARIMA model. In *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems (ITSC '09)*, 2009.

[37] C. Quek, M. Pasquier, and B. Lim. POP-TRAFFIC:A Novel Fuzzy Neural Approach to Road Traffic Analysis and Prediction. *IEEE Transactions on Inteliigent Transporation Systems*, 7(2):133–146, 2006.

[38] C. Robert. *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag, 1994.

[39] A. Statthopoulos and M. G. Karlaftis. A multivariate state space approach for urban traffic flow modeling and predicting. *Journal of Transportation Research Part C*, 11:121–135, 2003.

[40] C. Stutz and T.A. Runkler. Classification and Prediction of Road Traffic Using Application-Specific Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems*, 10(3):297–308, 2002.

[41] X. Sun, L. Munoz, and R. Horowitz. Mixture Kalman filter based highway congestion mode and vehicle density estimator and its application. In *Proceedings of the 2004 American Control Conference*, pages 2098–2103, Boston, MA, 2004.

[42] G.J. Szekely and M.L. Rizzo. Hierarchical clustering via joint between-within distances:extending ward's minimum variance method. *Journal of Classification*, vol.22:151–183, 2005.

[43] A. Thiagarajan, L. Sivalingam, K. LaCurts, S. Toledo, J. Eriksson, S. Madden, and H. Balakrishnan. VTrack: Accurate, Energy-Aware Traffic Delay Estimation Using Mobile Phones. In *7th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Berkeley, CA, November 2009.

[44] Y. Wang and M. Papageorgiou. Real-time freeway traffic state estimation based on extended kalman filter: a general approach. *Transportation Research Part B*, 39:141–167, 2005.

[45] D. Work, S. Blandin, O. Tossavainen, B. Piccoli, and A. Bayen. A traffic model for velocity data assimilation. *Applied Research Mathematics eXpress (ARMX)*, April 2010.

[46] H. Yin, S.C. Wong, J. Xu, and C.K. Wong. Urban traffic flow prediction using a fuzzy-neural approach. *Transportation Research Part C: Emerging Technologies*, 10(2):85–98, 2002.